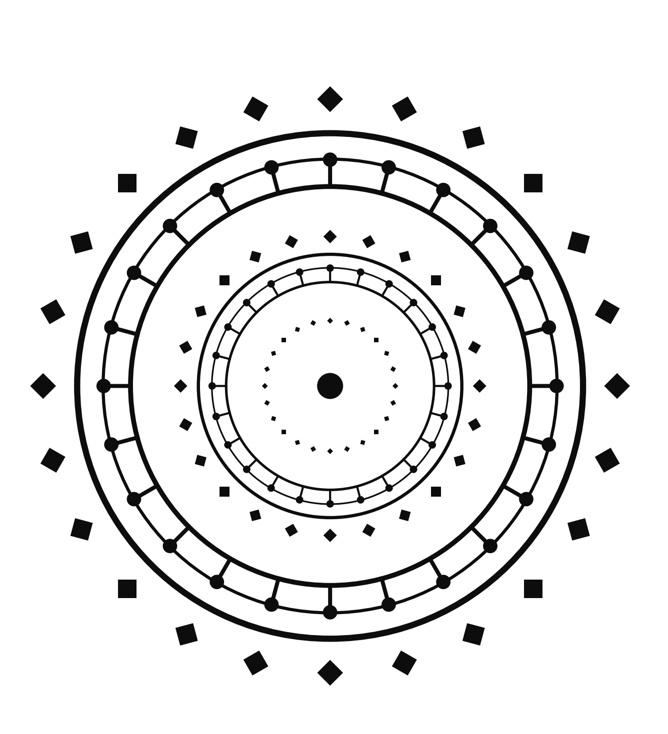
UNIVERSITE CADI AYYAD

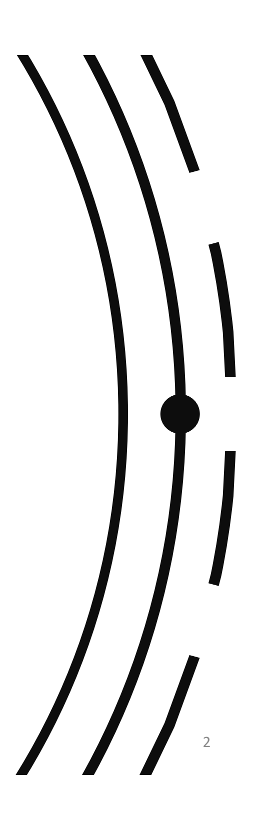
ECOLE NATIONALE DE COMMERCE ET DE GESTION
Marrakech

La statistique appliquée

Pr. Mourad TOUNSI
mouradetounsi@hotmail.com

1





2

PLAN

Introduction Générale: Omniprésence de la statistique comme science

Chapitre 1 : La description des données

Les statistiques sont un inventaire de techniques et de procédures qui permettent d'organiser et de faire le sommaire d'une masse d'informations afin d'en dégager des conclusions utiles à la compréhension d'un phénomène.

1. La description et l'inférence en statistique

2. L'organisation d'une banque de données pour l'analyse statistique

3. Les variables, dépendantes et indépendantes (endogènes, exogènes)

3. Les échelles de mesure

1. La description et l'inférence en statistique

- Les statistiques descriptives font le **sommaire** et simplifient l'information dans le but de la clarifier et de révéler ses **tendances** lourdes.
- **L'inférence** statistique est une série de procédures qui se servent de ces descriptions pour tirer des **conclusions** plus **générales** sur le phénomène à l'étude.
- Les domaines **d'application**: La sociologie, la psychologie, cognitivistes, marketing, gouvernement, écologie, finance, qualité, l'art, les affaires, …

5

2. L'organisation d'une banque de données pour l'analyse statistique

- **Organisation des banques de données**
  - **Logiciels**: Word ou Excel, ou
  - **Programmes** d'analyse statistique spécialisés tels que SPSS ou SAS (Statistical Analyse System 1976).
  - La règle : *chaque rangée définit un sujet différent et chaque colonne, une variable différente.*
- **Importation** dans un logiciel d'analyse statistique (SPSS: Statistical Pacadge for Sociales Sciences 1968-2017).

6

3. Les variables

- Une **caractéristique** que l'on **mesure** (soumise à des analyses):
- On l'appelle ***variable*** parce que les sujets d'analyse peuvent lui attribuer des valeurs différentes (l'âge, le sexe, le quotient intellectuel (QI) et la condition sociale)
  - Le QI est une variable parce qu'il peut être **différent** selon les personnes.
  - L'anxiété est une variable puisque certains peuvent être très anxieux, d'autres très calmes et d'autres encore peuvent se situer quelque part entre ces deux extrêmes.
  - Le genre – homme versus femme – est lui aussi une variable.
  - Si la variable ne peut prendre qu'une valeur unique, cette variable devient une constante. lorsque nous mesurons le degré de sociabilité des femmes, le sexe, qui est habituellement une variable, devient une constante (toutes les personnes mesurées étant des femmes).

4. Les échelles de mesure

- **Pourquoi on mesure les phénomènes?**
- **Une échelle de mesure ?**
- **Le sens de la différence entre les échelles?**

4. Les échelles de mesure

- Fournir une **valeur numérique** qui indique la **position** de **l'observation** sur la ou les **variables**.

- Les variables peuvent contenir **différents types d'informations**. Nous appelons le type d'informations l'« échelle » de mesure. (poids, course des chevaux).

- La signification des valeurs numériques que nous attribuons aux différents types de variables n'est pas toujours la même : (1 % à un « examen » ; 1er de classe; 1 pour les variable dichotomique binaires)

- 4 types d'échelles de mesure : **nominale, ordinale, à intervalles et de rapport**.

- Il est important de reconnaître l'échelle de mesure de chaque variable, car les **procédures statistiques** utilisables en dépendent.

9

4. Les échelles de mesure
Les variables (échelles) **nominales**

- Elle implique un simple **groupement** des observations en catégories **qualitatives** identifiées par un symbole (souvent une étiquette, tel « Homme » et « Femme » pour identifier le sexe).

- La seule opération mathématique possible avec cette échelle est de compter le nombre d'éléments (les effectifs) dans chacune des catégories (parfois nommées des classes), qu'on appelle aussi la **fréquence observée** ou plus simplement, la **fréquence**.

- **Quiz rapide**
  Vous devez coder la couleur des yeux de 1 000 personnes. Vous établissez les catégories « bleus = 1 », « bruns = 2 » et « verts = 3 ». Une personne a un œil bleu et l'autre vert. Comment allez-vous coder les yeux de cet individu ?

10

4. Les échelles de mesure
Les variables (échelles) **ordinales**

- L'échelle **ordinale** est **similaire** à l'échelle **nominale** exceptée qu'elle permet d'établir une relation **d'ordre** entre les éléments d'un ensemble, sans toutefois être capable d'évaluer de façon quantitative la distance qui les sépare.

  - Y-a-t-il un ordre entre les femmes et les hommes?
  - Notes scolaires.

- Une échelle **ordinale** représente des **rangs**. Avec cette échelle de mesure, on peut calculer des fréquences, mais aussi des moyennes et d'autres statistiques: L**a moyenne** doit être comprise comme le **rang moyen**.

11

4. Les échelles de mesure
Les variables (échelles) **à intervalles**

- L'échelle **relative** (encore appelé l'échelle à intervalles) définit **numériquement** les **intervalles** entre les données.

- Elle possède une **unité** de mesure **arbitraire** mais **constante**. Cependant, le zéro sur ces échelles est défini de façon arbitraire.

  - La température exprimée en Celsius. Zéro Celsius est un point arbitraire qui a été choisi par convention (les échelles Fahrenheit et Celsius n'ont pas le même zéro)
  - Cette échelle de mesure ne permet pas d'affirmer que de l'eau à 10 Celsius est deux fois plus chaude que de l'eau à 5 Celsius.

12

4. Les échelles de mesure

Les variables (échelles) **de rapport (absolue)**

- Elle implique que la **distance** entre deux unités est la **même** tout au long de l'échelle (tout comme dans l'échelle relative) mais aussi que le zéro existe (autrement que par un choix arbitraire).
- En plus de permettre de quantifier la différence entre deux éléments, elle permet aussi de calculer des rapports entre deux mesures.
  - Par exemple, une distance de 4 mètres est belle et bien le **double** d'une distance de 2 mètres.
  - La température en **Kelvin**. De l'eau à 300 Kelvin est deux fois plus chaude que de l'eau à 150 Kelvin en ce sens que l'on peut en extraire deux fois plus d'énergie cinétique.

13

4. Les échelles de mesure

Les relations entre les diverses échelles de mesure

- Les quatre types d'échelle ont été présentés dans un **ordre ascendant de précision.**
- Dans la suite, nous nommerons échelles de **type I** les échelles nominale et ordinale et échelles de **type II** les échelles relative et absolue.
- Il est toujours possible de passer d'une échelle d'un niveau donné à une échelle moins précise; l'inverse n'est cependant pas possible.
- **Quiz récapitulatif**
  Voici les résultats obtenus à un examen de statistique par trois étudiants :
  Marie = 90 %, Paul = 71 %, Julie = 70 %. Tirez les conclusions nominales, ordinales, à intervalles et de rapport pour ces trois observations.

14

4. Les échelles de mesure

Les relations entre les diverses échelles de mesure

TABLEAU 1 Comparaison des échelles de mesure					
Échelles de mesure		***Catégorie***	***Ordre***	***Différence relative***	***Différence absolue***
Type I	Nominale	✓			
	Ordinale	✓	✓		
Type II	À Intervalles	✓	✓	✓	
	De Rapport	✓	✓	✓	✓

15

4. Les échelles de mesure

Les relations entre les diverses échelles de mesure

Type 1

- a) ***Échelle nominale*** *:* Marie, Paul et Julie ont obtenu trois valeurs différentes.
- b) ***Échelle ordinale*** *:* Marie est arrivée première, Paul deuxième et Julie troisième (ou dernière).

Type 2

- c) ***Échelle à intervalles*** *:* Marie a beaucoup mieux réussi que Paul, qui n'a que légèrement mieux réussi que Julie.
- d) ***Échelle de rapport*** *:* la performance de Marie est de (90 – 71)/71 = 26,7 % meilleure que celle de Paul, et la performance de Paul est de (71 – 70)/70 = 1,4 % supérieure à celle de Julie.

16

CHAPITRE 2
LA DISTRIBUTION DES DONNÉES

1. La distribution simple des données
2. La distribution groupée des données
3. La distribution relative des données
   - La distribution cumulative : proportions et pourcentages
4. Les représentations graphiques de la distribution des données
   - Le graphique des histogrammes
   - Le polygone des effectifs
5. Les formes de distribution
   - La distribution unimodale La distribution bimodale (ou multimodale)
   - La distribution symétrique
   - La distribution asymétrique
   - Le degré d'aplatissement : leptocurtique Mésocurtique (normale) et platycurtique
6. La distribution des fréquences : voir le dossier 2 des TD.

17

CHAPITRE 2
LA DISTRIBUTION DES DONNÉES

- La statistique consiste à **réduire** une grande quantité d'informations à une expression plus **simple**, afin d'en tirer des renseignements **utiles**.
- **Comment** établir et représenter la distribution des données, numériquement et visuellement, à l'aide de graphiques?
- Comment **décrire** une information?
- Le point de départ de ce processus de simplification consiste à simplement **recenser** (compter) le nombre **d'observations** qui appartiennent à chaque **valeur** d'une **variable**.

18

1. La distribution simple des données

- Une **fréquence** (un effectif) est simplement le **décompte** du nombre d'observations ayant obtenu une certaine valeur.

- Elle organise les informations que contient la banque de données en regroupant ensemble celles qui sont **identiques** et permet ainsi d'en **réduire** le nombre.

- L'utilisation de la distribution **simple** des effectifs est tout à fait appropriée aux sondages sur les intentions de vote, dont on trouve les résultats dans les journaux.
  - Présentés sous forme de tableaux, ces résultats indiquent le nombre ou (plus généralement) le pourcentage des répondants qui se disent prêts à voter pour l'un ou l'autre des partis politiques. Puisque le nombre de partis politiques est relativement restreint, l'utilisation de la distribution simple représente une technique très efficace pour saisir rapidement le degré de popularité de chacun des partis.

2. La distribution **groupée** des données

- Les statistiques descriptives servent à **réduire** la **masse** d'informations afin de pouvoir s'en faire une idée **globale**.

- La construction d'une distribution **simple** des effectifs n'est pas toujours la manière la plus **pratique** pour faire le **sommaire** d'une banque de données.

- Lorsque les **valeurs** différentes sont **nombreuses**, la description de la variable devient **très détaillée**, ce qui **complexifie** l'interprétation que l'on peut en faire. Il est préférable de simplifier et de **réduire** davantage la banque de données.

- La distribution groupée des **fréquences** contiendra moins de catégories que la distribution simple

- La **simplification** de la banque de données augmente certes la **clarté** de l'information (facile d'en faire une **interprétation**) en **sacrifiant** des détails.

- En général, pour avoir une idée globale d'un ensemble de données, celui-ci ne doit pas contenir plus d'une **vingtaine de catégories**. Il nous faut donc réduire le nombre de catégories dans la variable.

2. La distribution **groupée** des données

- La façon de créer une distribution groupée des effectifs est très simple :
  1. On décide d'abord du nombre de catégories que l'on veut. Généralement, entre 10 et 20 catégories. Mais cette règle n'est pas coulée dans le béton. Pour certaines applications, il est approprié d'en créer plus de 20 ou moins de 10.
  2. Ensuite, on calcule la différence entre la plus petite et la plus grande valeur dans la distribution (cette différence, l'*étendue de la distribution*, est une statistique de base qui est décrite au chapitre 3).
  3. Enfin, on divise cette différence par le nombre de catégories. Le résultat obtenu indique la taille de chaque intervalle.
- Prenons les salaires des joueurs de la Real Madrid et établissons une distribution groupée des effectifs pour 10 intervalles.
  1. La différence entre le salaire le plus élevé et le plus bas est de 10 835 000 $ (11 000 000-165 000 $).
  2. Puisque nous désirons établir les effectifs pour 10 catégories de salaires, nous divisons l'étendue des salaires (10 835 000 $) par 10, et ainsi chaque intervalle regroupera les salaires en tranches de 1 083 500 $.
  3. Nous pouvons maintenant construire nos intervalles et établir la distribution groupée des données : la première catégorie compte le nombre de joueurs ayant un salaire situé entre 165 000 et 1 248 500 $ (165 000 $ + 1 083 500 $ = 1 248 500 $) et la deuxième inclut tous les salaires entre 1 248 501 et 2 332 000 $. Le dernier intervalle comprend tous les salaires entre 9 916 501 et 11 000 000 $.

21

2. La distribution **groupée** des données

Prenons les salaires des joueurs de la Real Madrid et établissons une distribution groupée des effectifs pour 10 intervalles.

1. La différence entre le salaire le plus élevé et le plus bas est de 10 835 000 $ (11 000 000-165 000 $).

2. Puisque nous désirons établir les effectifs pour 10 catégories de salaires, nous divisons l'étendue des salaires (10 835 000 $) par 10, et ainsi chaque intervalle regroupera les salaires en tranches de 1 083 500 $.

3. Nous pouvons maintenant construire nos intervalles et établir la distribution groupée des données : la première catégorie compte le nombre de joueurs ayant un salaire situé entre 165 000 et 1 248 500 $ (165 000 $ + 1 083 500 $ = 1 248 500 $) et la deuxième inclut tous les salaires entre 1 248 501 et 2 332 000 $. Le dernier intervalle comprend tous les salaires entre 9 916 501 et 11 000 000 $.

22

2. La distribution **groupée** des données

- La taille de l'intervalle créé par cette façon de faire produit un chiffre peu usuel (1 083 500 $). Or, il est généralement préférable d'arrondir la taille des intervalles. Ainsi, au lieu d'utiliser un intervalle de 1 083 500 $, il est plus commode de choisir un intervalle de 1 100 000 $.
- Donc, le premier intervalle comprend les salaires se situant entre 0 et 1 100 000 $ inclusivement, le deuxième intervalle, les salaires supérieurs à 1 100 000 $ et égaux ou inférieurs à 2 200 000 $, le troisième intervalle, les salaires supérieurs à 2 200 000 $ et égaux ou inférieurs à 3 300 000 $, etc. Le Tableau 2.1 montre les effectifs groupés pour les salaires des joueurs de la LNH. On peut y remarquer deux aspects importants :
  - Chaque salaire appartient à une seule catégorie.
  - Tous les salaires sont catégorisés.

Travail à faire: Etablir une distribution groupée des effectifs pour 20 intervalles.

2. La distribution **groupée** des données

TABLEAU 2
Distribution des données pour les salaires des joueurs du REAL MADRID, 2002-2003, avec intervalle de 1 100 000 $

Catégorie de salaires (intervalle) en M $	*Fréquence*	*Pourcentage (proportion)*	*Pourcentage cumulatif*
Plus de 0 à 1,1	374	55,1 % (0,551)	55,1 %
Plus de 1,1 à 2,2	148	21,8 % (0,218)	76,9 %
Plus de 2,2 à 3,3	76	11,2 % (0,112)	88,1 %
Plus de 3,3 à 4,4	30	4,4 % (0,044)	92,5 %
Plus de 4,4 à 5,5	20	2,9 % (0,029)	95,4 %
Plus de 5,5 à 6,6	9	1,3 % (0,013)	96,8 %
Plus de 6,6 à 7,7	5	0,7 % (0,007)	97,5 %
Plus de 7,7 à 8,8	5	0,7 % (0,007)	98,2 %
Plus de 8,8 à 9,9	7	1,0 % (0,01)	99,3 %
Plus de 9,9 à 11	5	0,7 % (0,007)	100,0 %
TOTAL	**679**	**100,0 % (1,0)**	

2. *La distribution groupée des données :* **sommaire des étapes**

La construction d'une distribution **groupée** des données: **trois règles** fondamentales.

1. Les intervalles définissant les catégories doivent être établis de manière à ce que **chaque observation** soit classée dans une **seule catégorie**.

2. Les **catégories** doivent être de **taille identique**. Elles respectent toutes la même étendue de valeurs de la variable.

1. Les **catégories** doivent être choisies de manière à couvrir toutes les **valeurs** possibles.

3. La distribution **relative** des données (*proportion*)

Le Tableau 2 est utile pour faire une représentation des salaires des joueurs de hockey:

- 374 joueurs sont payés 1 100 000 $ ou moins, tandis que seulement 5 gagnent 9 900 000 $ ou plus. Il va sans dire qu'un salaire aux alentours de 1 000 000 $ est plus habituel dans la REAL MADRID qu'un salaire de 10 000 000 $.

- Pour mieux comprendre ces effectifs, il est souvent pratique d'exprimer, pour chaque valeur ou catégorie de valeurs, la fréquence des observations qui s'y trouvent relativement au nombre total d'observations. Cette distribution prend un nom différent. On l'appelle *distribution relative des effectifs,* car la fréquence des observations pour chaque valeur exprime le nombre d'observations dans chaque valeur *relative* (par rapport) au nombre total d'observations. Nous pouvons exprimer ce rapport en *proportion* ou en *pourcentage*.

- La proportion indique la fréquence des observations se trouvant dans chaque intervalle relatif au nombre total d'observations. Le calcul de la proportion est facile : il s'agit simplement de diviser la fréquence obtenue pour chaque intervalle (fi) par le nombre total d'observations (*N*) : Proportion = fi/*N. (formule 1)*

3.1. La distribution **cumulative** : proportions et pourcentages

Revenir au tableau

Exemple:

Diagramme de Pareto et la règle 20 / 80

4. Les représentations graphiques de la distribution des données

4.1. Le graphique des histogrammes

4.2. Le polygone des effectifs

4.2. Le polygone des effectifs

- Lorsqu'on travaille avec des variables à intervalles ou de rapport, on peut aussi remplacer l'histogramme par une ligne liant les fréquences (un *graphique des polygones):*
  - Les polygones des effectifs sont souvent plus lisibles que les histogrammes
  - Ils sont pratiques lorsque utilisés pour décrire des distributions de fréquences relatives.
- La construction d'un polygone des fréquences est très simple.
  - Lorsqu'on travaille avec des distributions simples, il s'agit de mettre un point sur le graphique se rapportant à la fréquence de chaque valeur de la variable, et de relier ensuite chacun de ces points par une ligne.
  - Lorsqu'on travaille avec des distributions groupées, on met le point à la valeur qui définit le centre de l'intervalle. Pour le polygone des salaires des joueurs de la REAL DE MADRID, le point qui décrit la première catégorie (0-1 100 000 $) est situé visuellement
    au centre de l'intervalle (550 000 $).
    Le polygone des fréquences utilise la même information que l'histogramme, et ces formes graphiques proviennent toutes deux de la distribution. L'avantage du polygone sur l'histogramme est qu'il produit un graphique visuellement plus simple. Si on étudie la Figure 2.2, on voit très
    bien que la fréquence des salaires plus élevés chute de façon marquante.

29

4.2. Le polygone des effectifs

- Lorsqu'on travaille avec des variables à intervalles ou de rapport, on peut aussi remplacer l'histogramme par une ligne liant les fréquences ; on appelle le résultat un *graphique des polygones.*

- Les polygones des effectifs sont souvent plus lisibles que les histogrammes et, comme nous le verrons plus loin, ils sont pratiques lorsque utilisés pour décrire des distributions de fréquences relatives.

- La construction d'un polygone des fréquences est très simple. Lorsqu'on travaille avec des distributions simples, il s'agit de mettre un point sur le graphique se rapportant à la fréquence de chaque valeur de la variable, et de relier ensuite chacun de ces points par une ligne. Lorsqu'on travaille avec des distributions groupées, on met le point à la valeur qui définit le centre de l'intervalle.

30

5. Les formes de distribution

a. La Distribution Unimodale

b. La Distribution Bimodale (ou Multimodale)

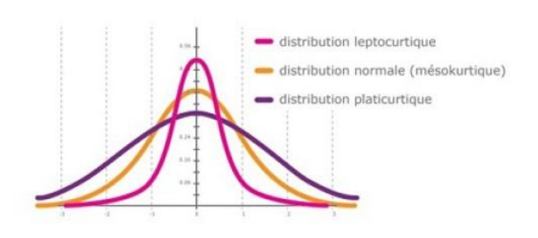
c. La Distribution Symétrique

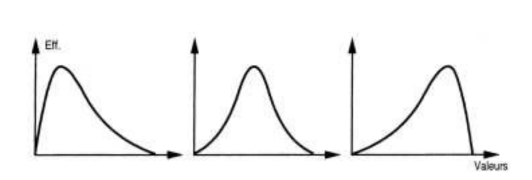
d. La Distribution Asymétrique

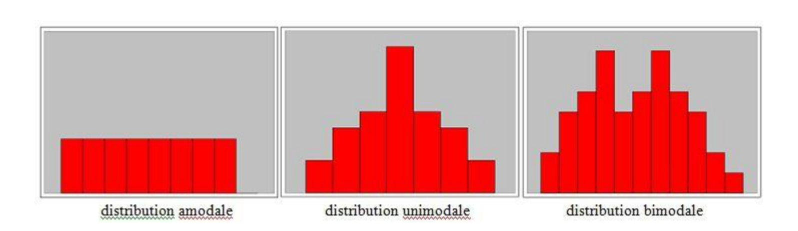
e. Le degré d'Aplatissement : Leptocurtique et Platycurtique

31

CHAPITRE 2
LA DISTRIBUTION DES DONNÉES

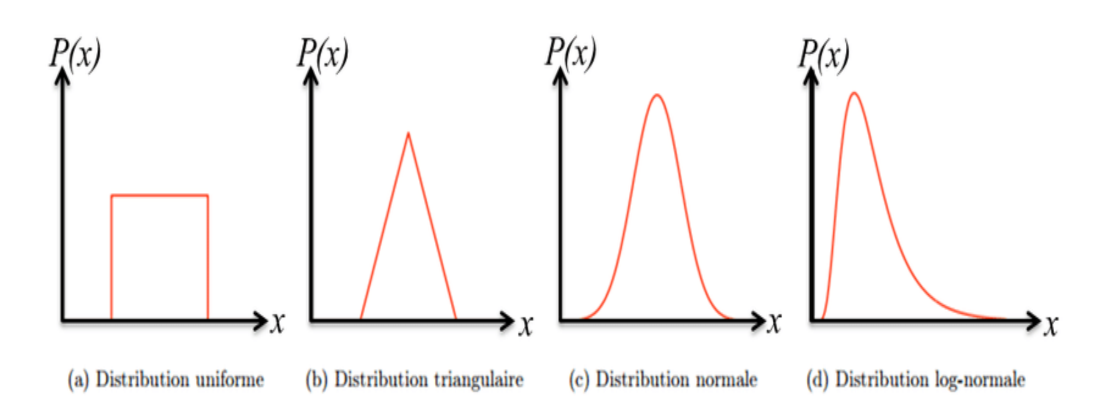






32

CHAPITRE 2
LA DISTRIBUTION DES DONNÉES



(a) Distribution uniforme (b) Distribution triangulaire (c) Distribution normale (d) Distribution log-normale

CHAPITRE 2
LA DISTRIBUTION DES DONNÉES

6. La distribution des fréquences : un exemple complet (voir dossier 2 - TD)